An Enhanced Topic Modeling Method in Educational Domain by Integrating LDA with Semantic

Ruofei Ding , Pucheng Huang, Shumin Chen, Jiale Zhang, Jingxiu Huang, Yunxiang Zheng✉

School of Educational Information Technology, South China Normal University, China

dr.zheng.scnu@hotmail.com

Abstract— **With the development of online courses, students' discussion texts in online forums and communication groups are increasing. Teachers can use these texts to monitor student learning so that they can adapt the pace of instruction accordingly. And textual topics, as the important information of the text, can be extracted from the text by topic modeling. Currently, a Latent Dirichlet Allocation (LDA) method has been used to identify the critical main topics discussed by students. However, LDA is based on word frequency and ignores semantic information. In this study, we propose a model for fusing semantic information into LDA. To verify the validity of our model, we collected two MOOC datasets for testing and conducted an ablation study using Silhouette Coefficient value and Calinski-Harabasz score as the criterion. The results show that our method is scientifically feasible and better than LDA in the field of educational topic modeling. Thus, our method is able to perform topic modeling more accurately compared to LDA. It can be used by teachers to automatically analyze large amounts of student discussion data to guide personalized learning paths.**

Keywords— **Topic Modeling, Online Discussion, Text Mining, Machine Learning, LDA**

I. Introduction

In today's world, a large number of texts containing student discussions can be found on online education platforms. These discussion texts serve as reflections of students' discussions on various topics, offering valuable insights for teachers to interpret learning outcomes. But analyzing textual data of student discussions without automation has become an almost impossible task. With the help of topic modeling, it is possible to automatically cluster different discussion texts together.

Latent Dirichlet Allocation (LDA), one of the existing topic models, can determine the main topic of students' discourse. LDA was created based on Bayesian probability and is widely used in the field of education because of its reasonable inference processes and hypothetical designs, combined with its requirement for only a few parameters to configure.

LDA is always used to summarize the keywords and the trends of topics in recent research in education, and is also used to improve the quality of personalized resource recommendations.

Odden used LDA to perform topic analysis on research published in the Physics Education Research Conference Proceedings from 2001 to 2018[1]. They aimed to analyze the trends of research topics and identify the topics which have received consistent attention. Through this research, they supposed that LDA was expected to help in educational research literature, referring that the analysis was "quantitative, independent, and replicable". Gurcan Fatih et al. used N-gram modeling together with LDA to study e-learning articles related to the COVID-19, and found the trends of research and popular research issues during that time[2].

Due to the capability of LDA to assist in constructing interest models for learners and computing user preferences, some researchers utilized LDA to optimize personalized resource recommendations in online education. Lin Qi et al. achieved such optimization using LDA[3]. Peng Jiang et al. combined LDA with Artificial Neural Networks for intelligent user recommendation of online video courses[4]. Wei Kuang et al. also utilized LDA to construct user interest models and proposed a resource recommendation method for e-learning systems[5].

The widespread application of LDA models in the online education domain can be attributed to its capacity to enhance the quality of feedback loops and its advantages over some other methods. For example, Chai et al.[6] introduced a method that uses LDA to detect topics in online course feedback. The method can present the topics of feedback to the teachers in the form of word clouds and analyze the relationship between the feedback and various factors such as students' grade, satisfaction and learning outcomes. Deepak and Shobha[7] used LDA to address the issue of identifying students who fail to complete assigned tasks within the given time in an online learning system. They employed LDA to cluster texts and learners, and the results showed that it achieved significant performance compared to other existing algorithms.

However, this does not imply that LDA is the optimal solution. There are still certain limitations in LDA. Li et al. pointed out that the LDA model fails to use semantic information to enhance feature representation, which may impact the results of semantic analysis. Grootendorst identified a limitation of the LDA model. It ignores the semantic between words due to its use of a bag-of-words representation, which leads to the result that the texts may not be represented accurately[8]. Tajbakhsh, Mir Saman also indicated that LDA disregards the semantic relationships between words in short text clustering[9].

To solve the problem of LDA lacking semantic information, it may be necessary to combine other methods with LDA to further represent the semantics, thus improving the performance of the model. This viewpoint was also shown in the article by Ekinci Ekin et al. They argued that traditional topic models have a significant limitation in which they cannot capture topics related to semantics. Furthermore, they emphasized the crucial role of semantic inference in topic modeling[10]. There are current studies indicating that semantic information can indeed affect the effectiveness of topic modeling. Grootendorst[8] found that it has a better performance in coherence of the result than LDA when using BERTopic for dynamic topic modeling. The topic coherence score is also significantly higher when using Word2vec combined with LSA instead of PLSA[11]. In their analysis of online discourse related to the Hong Kong extradition bill incident, Xu[12] found that there is a better topic relevance when combining LDA with BERT, with a 35.7% enhancement compared to using LDA only.

In text clustering analysis, Li[13] used Word2Vec in combination with LDA for topic modeling and clustering analysis of academic article abstracts. The results showed that their approach achieved approximately a 9% higher accuracy compared to using LDA only. Similarly, George and Sumathy[14] used BERT in combination with LDA for topic modeling and clustering analysis of the open dataset CORD-19, finding that their approach performed at least 10% better than using LDA only.

It can be seen that incorporating semantic information into topic modeling can significantly enhance its performance, enabling researchers to conduct more in-depth analysis of the results of topic modeling. In the current education domain, there is still limited research on combining semantic information with LDA for topic modeling and text clustering analysis. Our research aims to propose a semantic-fusioned LDA topic modeling algorithm for topic modeling and clustering of educational texts.

II. Methodology

A. Data Preparation

To prove the stability and reliability of our method, we collected two datasets (DATASET 1 and DATASET 2) for the years 2018 and 2022 from the course "Instructional Design Principles and Methods" on the China University MOOC Platform. It was a 15-week introductory Educational technology course. It provided learners with course materials, lecture videos, reading materials, and test questions, as well as forums to support peer interactions. Both datasets were obtained from the interactive forum where students and teachers engaged in discussions. Each row of the two datasets contains the question, the student's account and the student's answer. DATASET 1 consists of eight topics with 1397 rows of raw text. DATASET 2 consists of five topics with 758 rows of raw text. We pre-processed DATASET 1 and DATASET 2 by removing duplicates and blacks, and ended up with 1343 left in DATASET 1 and 703 left in DATASET 2.

B. Text representation with semantic

Text should first be transformed into a suitable representation before it can be used as data[15]. The representation determines the effectiveness in natural language processing(NLP) tasks. In the early days, researchers commonly used one-hot coding and TF-IDF coding, but both of them could only represent limited information. With the development of deep learning, the representation of text has shifted from discrete words to continuous vectors.

Continuous n-dimensional vectors can capture semantics. Word2vec[16] and Glove[17] are pre-trained word embedding models that are frequently used to convey semantics through a continuous vector. However, they are static in capturing lexical dimensions and neglect variations of semantic, so they can't represent long-term dependencies between words. In order to better represent semantics, ELMo[18] and BERT[19] have emerged. They generate dynamic word vectors for all words based on context. But ELMo employs a recurrent neural network(RNN), thus ELMo has shortcomings in learning long-term dependencies. In contrast, BERT is based on multi-head attention. So BERT is good at resolving long-term dependencies of text and can therefore represent semantic information of longer texts.

BERT was trained on the BooksCorpus dataset (800 million words) and text passages from the English Wikipedia. BERT can be used on unannotated data directly from a pre-trained model, or it can be fine-tuned for task-specific data. The most common variants of BERT are Roberta[20], DistilBERT[21], XLNet[22], ALBERT[23] and ERNIE 2.0[24]. Among them, ERNIE adapted the MASK disambiguation technique and was trained on a Chinese corpus. Therefore, ERNIE has significant improvements in Chinese NLP tasks. In this paper, we use the pre-trained ERNIE to generate embedding features for each text.

C. Proposed methodology

In this paper, we present a method (Figure 1) to LDA topic modeling that incorporates semantics in the educational domain. First of all, we used LDA for topic modeling in DATASET 1 and DATASET 2, then we obtained probability vectors (PVs) of the text belonging to each topic. Next, considering ERNIE's strengths in Chinese text， we use it to obtain sentence embedding (SEs) containing semantic information. Then, we combined PVs and SEs to obtain "Topic - Semantics" vectors (TSVs), a type of non-linear data. Later, we used ISOMAP[25] to perform dimensionality reduction on the TSVs and obtain DTSVs. Finally, we used the K-means algorithm to cluster the DTSVs. While K-means chooses centroids randomly before clustering, once the centroids are poorly chosen, it may lead to unsatisfactory clustering results. Thus, we used the Particle Swarm Optimisation (PSO) algorithm to optimize K-means.

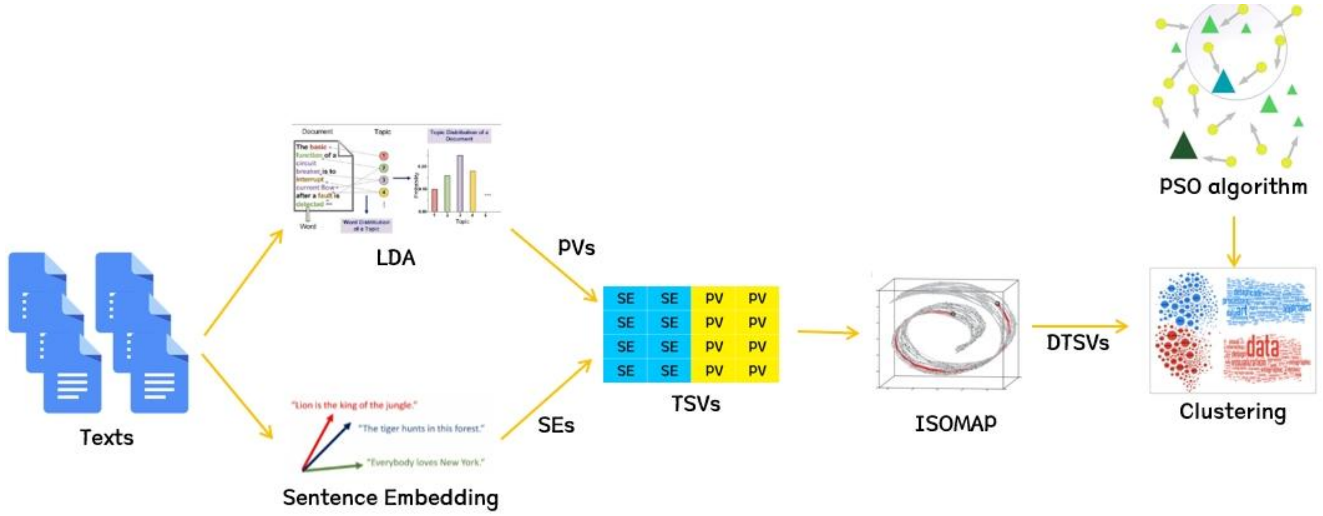

Texts
LDA
"Lion is the king of the jungle."
"The tiger hunts in this forest."
"Everybody loves New York."
Sentence Embedding
PVs
SEs
SE SE PV PV
TSVs
ISOMAP
DTSVs
PSO algorithm
Clustering

Figure 1. A topic modeling method for LDA incorporating semantics in the educational domain

PSO was invented by James Kennedy and Russell Eberhart inspired by the regularity of the foraging behavior of birds[26]. The algorithm works by initializing a flock of birds randomly over the searching space, where each bird is referred to as a ‘‘particle’’.

Consider that a set of ‘‘particles’’ fly with a certain velocity algorithm and move to find the global best position in an iterative process. At each iteration of the algorithm, the velocity vector for each particle is modified based on three parameters: the particle momentum（The current speed of the particle）, the best position reached by the particle and that of all particles up to the current stage.

The positions and velocities of the particles are calculated using equation (1) and equation (2).

$$x_i = x_i + v_i \qquad (1)$$

$$v_i = w \times v_i + c_1 \times rand() \times (pbest_i - x_i) + c_2 \times rand() \times (gbest_i - x_i) \qquad (2)$$

Where i is the hyperparameter representing the total number of particles. x_i is the current position of the particle. v_i is the directed velocity of the particle, representing the memory term（momentum). w is the learning rate, indicating the efficiency of the particle swarm learning after each iteration. rand() is a random number between (0,1). $pbest_i$ represents the current local searched optimum position searched by the particle. $gbest_i$ represents the current searched optimum position of the swarm. $c_1 \times rand() \times (pbest_i - x_i)$ and $c_2 \times rand() \times (gbest_i - x_i)$ represent the particle pi 's cognitive and the global cognitive of all particles respectively.

PSO continuously adjusts the distance between the initial centroid and the global optimal centroid by continuous iteration. Using the outcome of the PSO as the initial centroids for K-means can effectively improve the result of K-means as these centroids are close to the global optimal centroids. Equation (3) is used to evaluate the clustering effect in each iteration.

$$F(x) = \sum_{i=1}^{k} (C_{labels=i} - centroids_i)^2 \qquad (3)$$

Where $C_{labels=i}$ represents the set of vectors that are labeled with i after the KMEANS clustering in the current state. The value of F(x) represents the total sum of squared distances between each label and the vectors belonging to that label. A smaller value of F(x) indicates a better clustering result for K-means, and indicates that the current position of the particle is more optimal.

D. Evaluation Metrics

To test our model, the Calinski-Harabasz (CH) Score and the Silhouette Coefficient (SC) were used as criteria to evaluate the result.

The CH score is the ratio of inter-cluster distance to intra-cluster distance and is defined as follows:

$$CH = \frac{(\sum_{k=1}^{K} n_k ||\mu_{c_k} - \mu||_2^2)(N-K)}{(\sum_{k=1}^{K} \sum_{i=1}^{n_k} ||x_i - \mu_{c_k}||_2^2)(K-1)} \qquad (4)$$

Where is the number of members in cluster, is the capacity of the dataset, indicates the number of clusters and represents the centroid of the dataset. The range of score is (0, +∞). The higher the value of the CH index is, the better the clustering validity is, that is, clusters are primely separated from each other and are distinctly preferable.

The SC evaluates the effect of clustering through cohesion and separation.The SC defined as:

$$SC = \frac{\sum_{i=1}^{N} \frac{b-a}{max(a,b)}}{N} \qquad (5)$$

Where a is the average distance from this sample to other samples in the same cluster, b is the average distance from this sample to all samples in the nearest neighboring cluster, N is the number of clustered. The range of SC is [-1, 1]. If the SC is close to -1, it indicates poor clustering and there are many samples that should be grouped in the neighboring cluster. If the SC is close to 0, it indicates that there are large areas of overlap between clusters. If the SC is close to 1, it indicates good clustering.

III. RESULT

To compare our method with LDA in terms of performance improvement, the number of a priori topics was adjusted to the number of topics in the original dataset. Specifically, in DATASET 1, we set the number of a priori topics for LDA to 8. In DATASET 2, the number of a priori topics is set to 5. Such a setup can better evaluate the performance of the LDA model and our proposed method on different datasets while ensuring fairness. Figure 2 and Figure 3 are word cloud results of our method for some topics examples in DATASET 1 and DATASET 2.In the three topic samples of DATASET 1, the main topics student concern about were "Instruction, Design", "Student, Instruction", and "Design, Curriculum". In the three topic samples of DATASET 2, they were "Analysis, Evaluation", "Instruction, Design", and "Training, Analysis". It can be seen from the results above that students focus on different topics when facing different topic samples, which leads to the discrepancy on the topics among these samples.

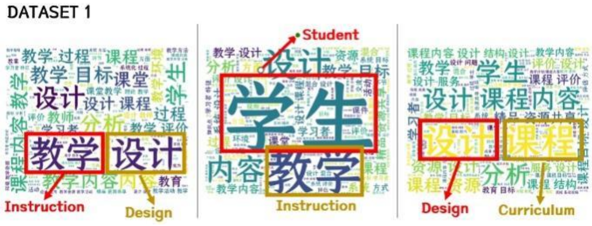


Figure 2. Some examples of topic word clouds from DataSet 1

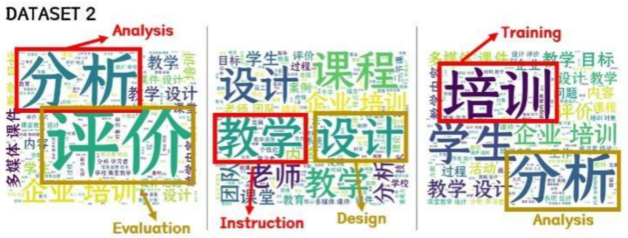


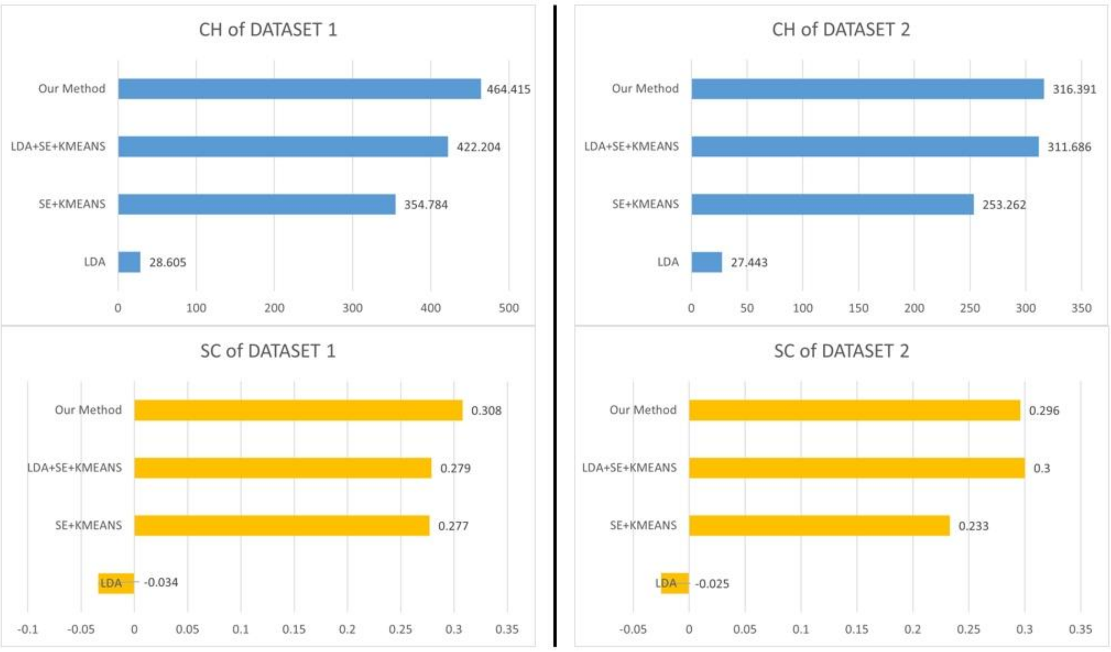
Figure 3. Some examples of topic word clouds from DataSet 2

To further analyze the scientific validity of our proposed method and the performance improvement of our method, we also conducted an ablation study using CH score and SC value. Figure 4 shows the results on two datasets, comparing the CH scores and SC values obtained by clustering using LDA, Sentence Embedding (SE)+KMEANS, LDA+SE+KMEANS, and our method, respectively.

We can see that our method shows superior performance in all models. Specifically, on DATASET 1, our model achieves the highest performance on both metrics. While on DATASET 2, compared to the next best performing LDA + SE + KMEANS model, our model has a higher CH score but a slightly lower SC value. We hypothesized that this may be due to the fact that the topics in DATASET 2 are very different from each other and the topics are not tightly structured internally. To verify our hypothesis, we visualized our data.

Figure 5 and Figure 6 show the clustering results of our method and other methods on DATASET 1 and DATASET 2. Clearly, the data and clustering results of DATASET 1 are closer than those of DATASET 2, which confirms our hypothesis.

Figure 4. CH SCORE AND SC VALUE OF TWO DATASETS WITH DIFFERENT METHODS



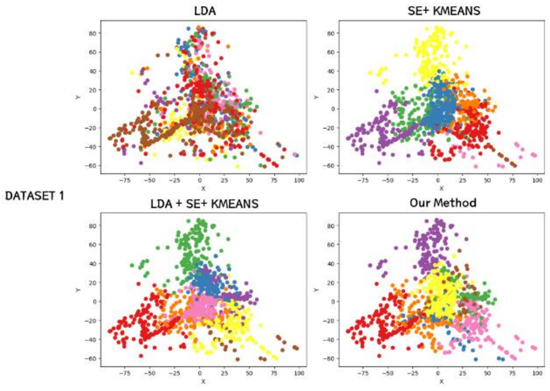


Figure 5. Visualization of clustering for DATASET 1

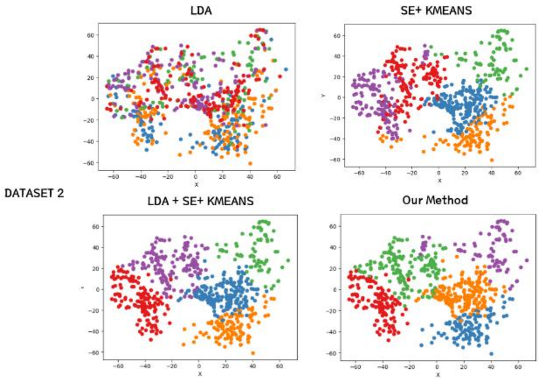


Figure 6. Visualization of clustering for DATASET 2

When comparing results on DATASET 1 and DATASET 2 by using our method, our model displays significantly greater metric differences on DATASET 1 when compared to the second-best LDA + SE + KMEANS method. Therefore, this result implies that our method is more effective in dealing with datasets with more clustered topic distributions. The PSO algorithm played a key role in achieving this result by optimizing the initial centroid selection of K-means to better handle densely distributed data. This also confirms the power of PSO for nonlinear optimization problems.

IV. Conclusion

In this study, we proposed a model for fusing semantic information with LDA. Specifically, we collected two MOOC datasets for testing and conducted an ablation study using SC value and CH score as the criterion. The results showed that our method is scientifically feasible and better than LDA in the field of educational topic modeling.

The innovation of our method is to incorporate semantic information into the LDA topic model and apply it to education. We validated the feasibility and effectiveness of the method in terms of performance. In future research, we will investigate whether the topic model incorporating semantic information can reflect students' cognition and analyze the results at a fine-grained level.

Acknowledgment

This research was funded by the National Natural Science Foundation of China: 62207015, and the Humanities and Social Sciences Youth Foundation of the Chinese Ministry of Education: 22YJC880021.

References

[1] T. O. B. Odden, A. Marin, and M. D. Caballero, “Thematic analysis of 18 years of physics education research conference proceedings using natural language processing,” *Phys. Rev. Phys. Educ. Res.*, vol. 16, no. 1, p. 010142, Jun. 2020, doi: 10.1103/PhysRevPhysEducRes.16.010142.

[2] F. Gurcan, G. G. M. Dalveren, and M. Derawi, “Covid-19 and E-Learning: An Exploratory Analysis of Research Topics and Interests in E-Learning During the Pandemic,” *IEEE Access*, vol. 10, pp. 123349–123357, 2022, doi: 10.1109/ACCESS.2022.3224034.

[3] Q. Lin, S. He, and Y. Deng, “Method of personalized educational resource recommendation based on LDA and learner’s behavior,” *International Journal of Electrical Engineering & Education*, p. 0020720920983511, Jan. 2021, doi: 10.1177/0020720920983511.

[4] P. Jiang, Y. Feng, C. Niu, and Y. Dai, “Study of intelligent recommendation for online video courses,” in *2021 IEEE 5th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC)*, Oct. 2021, pp. 1290–1294. doi: 10.1109/ITNEC52019.2021.9587262.

[5] W. Kuang, N. Luo, and Z. Sun, “Resource recommendation based on topic model for educational system,” in *2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference*, Aug. 2011, pp. 370–374. doi: 10.1109/ITAIC.2011.6030352.

[6] S. Unankard and W. Nadee, “Topic Detection for Online Course Feedback Using LDA,” in *Emerging Technologies for Education*, E. Popescu, T. Hao, T.-C. Hsu, H. Xie, M. Temperini, and W. Chen, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 133–142. doi: 10.1007/978-3-030-38778-5_16.

[7] N. A. Deepak and N. S. Shobha, “Analysis of Learner’s Behavior Using Latent Dirichlet Allocation in Online Learning Environment,” in *Computational Methods and Data Engineering*, V. Singh, V. K. Asari, S. Kumar, and R. B. Patel, Eds., in Advances in Intelligent Systems and Computing. Singapore: Springer, 2021, pp. 231–242. doi: 10.1007/978-981-15-7907-3_18.

[8] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure.” arXiv, Mar. 11, 2022. doi: 10.48550/arXiv.2203.05794.

[9] M. S. Tajbakhsh and J. Bagherzadeh, “Semantic knowledge LDA with topic vector for recommending hashtags: Twitter use case,” *Intelligent Data Analysis*, vol. 23, no. 3, pp. 609–622, Jan. 2019, doi: 10.3233/IDA-183998.

[10] E. EKİNCİ and S. OMURCA, “NET-LDA: a novel topic modeling method based on semantic document similarity,” *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 4, pp. 2244–2260, Jan. 2020, doi: 10.3906/elk-1912-62.

[11] S. Kim, H. Park, and J. Lee, “Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis,” *Expert Systems with Applications*, vol. 152, p. 113401, Aug. 2020, doi: 10.1016/j.eswa.2020.113401.

[12] X. Tan, M. Zhuang, X. Lu, and T. Mao, “An Analysis of the Emotional Evolution of Large-Scale Internet Public Opinion Events Based on the BERT-LDA Hybrid Model,” *IEEE Access*, vol. 9, pp. 15860–15871, 2021, doi: 10.1109/ACCESS.2021.3052566.

[13] C. Li *et al.*, “LDA Meets Word2Vec: A Novel Model for Academic Abstract Clustering,” in *Companion Proceedings of the The Web Conference 2018*, in WWW ’18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, Apr. 2018, pp. 1699–1706. doi: 10.1145/3184558.3191629.

[14] L. George and P. Sumathy, “An integrated clustering and BERT framework for improved topic modeling,” *Int. j. inf. tecnol.*, vol. 15, no. 4, pp. 2187–2195, Apr. 2023, doi: 10.1007/s41870-023-01268-w.

[15] K. Babić, S. Martinčić-Ipšić, and A. Meštrović, “Survey of Neural Text Representation Models,” *Information*, vol. 11, no. 11, Art. no. 11, Nov. 2020, doi: 10.3390/info11110511.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space.” arXiv, Sep. 06, 2013. Accessed: Sep. 28, 2023. [Online]. Available: http://arxiv.org/abs/1301.3781

[17] J. Pennington, R. Socher, and C. Manning, “GloVe: Global Vectors for Word Representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[18] M. E. Peters *et al.*, “Deep contextualized word representations.” arXiv, Mar. 22, 2018. Accessed: Sep. 28, 2023. [Online]. Available: http://arxiv.org/abs/1802.05365

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv, May 24, 2019. doi: 10.48550/arXiv.1810.04805.

[20] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” arXiv, Jul. 26, 2019. doi: 10.48550/arXiv.1907.11692.

[21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.” arXiv, Feb. 29, 2020. doi: 10.48550/arXiv.1910.01108.

[22] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019. Accessed: Sep. 28, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

[23] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.” arXiv, Feb. 08, 2020. doi: 10.48550/arXiv.1909.11942.

[24] Y. Sun *et al.*, “ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, Art. no. 05, Apr. 2020, doi: 10.1609/aaai.v34i05.6428.

[25] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, “A Global Geometric Framework for Nonlinear Dimensionality Reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000, doi: 10.1126/science.290.5500.2319.

[26] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in *Proceedings of ICNN’95 - International Conference on Neural Networks*, Nov. 1995, pp. 1942–1948 vol.4. doi: 10.1109/ICNN.1995.488968.